Predviđanje visine godišnjih prihoda na temelju popisnih podataka

V. I. Banić

P. Bratulić

L. Bundalo

Opis problema

- Demografske i ekonomske karakteristike
- Podatak o prihodima – dvije kategorije
- 14 atributa
- 48 842 instance
  - Train dana (32 561)
  - Test dana (16 281)

Dosadašnja istraživanja

- Usporedbe nekih algoritama
  - Naivni Bayes, Stablo odluke, NBTree
  - Naivni Bayes, Logistička regresija, Slučajne šume
    - Bez balansiranja
    - Random Oversampling
    - Random Undersampling
    - kombinacija
- Koristi se i algoritam K najbližih susjeda

Plan istraživanja

- Implementacija više klasifikacijskih algoritama
  - Logistička regresija, neuronske mreže, K najbližih susjeda, …
- Python paket sklearn
- Uspješnost: preciznost (accuracy)
- Usporedba različitih algoritama
- Usporedba s već poznatim rezultatima

Literatura

- [1] http://archive.ics.uci.edu/ml/datasets/adult(Zadnje pristupljeno 18.4.2019.) Originalni dataset i opis problema
- [2] http://robotics.stanford.edu/~ronnyk/nbtree.pdf(Zadnje pristupljeno 18.4.2019.) Rad u kojem je prvi put korišten promatrani skup podataka.
- [3] https://storage.googleapis.com/kaggle-forum-message-attachments/160002/5905/Paper%20on%20Machine%20Learning%20for%20Kaggle.pdf (Zadnje pristupljeno 18.4.2019.)